University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2015

Stuck and Frustrated or In Flow and Happy: Sensing Developers' Emotions and Progress

Müller, Sebastian C ; Fritz, Thomas

Abstract: Software developers working on change tasks commonly experience a broad range of emotions, ranging from happiness all the way to frustration and anger. Research, primarily in psychology, has shown that for certain kinds of tasks, emotions correlate with progress and that biometric measures, such as electro-dermal activity and electroencephalography data, might be used to distinguish between emotions. In our research, we are building on this work and investigate developers' emotions, progress and the use of biometric measures to classify them in the context of software change tasks. We conducted a lab study with 17 participants working on two change tasks each. Participants were wearing three biometric sensors and had to periodically assess their emotions and progress. The results show that the wide range of emotions experienced by developers is correlated with their progress on the change tasks. Our analysis also shows that we can build a classifier to distinguish between positive and negative emotions in 71.36% and between low and high progress in 67.70% of all cases. These results open up opportunities for improving a developer's productivity. For instance, one could use such a classifier for providing recommendations at opportune moments when a developer is stuck and making no progress.

DOI: https://doi.org/10.1109/ICSE.2015.334

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-108927
Conference or Workshop Item
Accepted Version

Originally published at:
Müller, Sebastian C; Fritz, Thomas (2015). Stuck and Frustrated or In Flow and Happy: Sensing Developers' Emotions and Progress. In: 37th International Conference on Software Engineering, Florence, Italy, 20 May 2015 - 22 May 2015, IEEE.
DOI: https://doi.org/10.1109/ICSE.2015.334

Stuck and Frustrated or In Flow and Happy: Sensing Developers' Emotions and Progress

Sebastian C. Müller, Thomas Fritz
Department of Informatics, University of Zurich, Switzerland
{smueller, fritz}@ifi.uzh.ch

***Abstract*—Software developers working on change tasks commonly experience a broad range of emotions, ranging from happiness all the way to frustration and anger. Research, primarily in psychology, has shown that for certain kinds of tasks, emotions correlate with progress and that biometric measures, such as electro-dermal activity and electroencephalography data, might be used to distinguish between emotions. In our research, we are building on this work and investigate developers' emotions, progress and the use of biometric measures to classify them in the context of software change tasks. We conducted a lab study with 17 participants working on two change tasks each. Participants were wearing three biometric sensors and had to periodically assess their emotions and progress. The results show that the wide range of emotions experienced by developers is correlated with their perceived progress on the change tasks. Our analysis also shows that we can build a classifier to distinguish between positive and negative emotions in 71.36% and between low and high progress in 67.70% of all cases. These results open up opportunities for improving a developer's productivity. For instance, one could use such a classifier for providing recommendations at opportune moments when a developer is stuck and making no progress.**

I. Introduction

Frustration, anger, happiness and enthusiasm are emotions that software developers frequently experience during their work [1]. These emotions are commonly intertwined with the progress one makes, such as experiencing positive emotions leading to more progress [2] or the state of being stuck and making no progress leading to frustration [3]. Research in psychology has already shown that there is a correlation between these two dimensions, the emotions and the progress people experience for certain kinds of tasks (*e.g.* [4]). To help ensure a developer's time is spent as productive as possible, an indicator for a developer's emotions could thus be used to prevent interruptions when a developer is "in flow", making a lot of progress and should not be disturbed, or to provide recommendations at opportune moments when the developer is getting frustrated and close to being stuck.

With the recent advances in biometric (*aka* psycho-physiological) sensor technology, an increasing amount of research in psychology has shown that a person's biometric features, such as skin temperature, facial expression or respiration rate, can be used to detect and distinguish between emotions (*e.g.* [5], [6]). Psychology research has also shown that biometric measures can be used to determine a flow or stuck state (*e.g.* [3], [7]). However, these studies are focused on small analytical tasks or physics exercises and do not provide any evidence on its applicability to software development tasks, in particular, given the complexity and emotions as well as cognitive skills these kinds of tasks stress in humans.

In software engineering, only little research has focused on developers' emotions and the use of biometric measures. For emotions, researchers have looked at the emotions that developers experience [1], [8], how they might affect productivity [9], [10], and whether one could use interaction logs to predict them [11], [12]. Using biometric sensors, in particular eye-tracking and fMRI, researchers have mainly studied how software developers comprehend code or use tools [13]–[15]. In a previous study, we looked at the use of biometric sensors to assess the difficulty of small code comprehension tasks [16].

In the research presented in this paper, we built upon existing work in software engineering and psychology and further investigate emotions and progress developers experience, as well as the use of biometric sensors to predict them in the context of change tasks. In particular, we are interested in the following three research questions:

RQ1: What is the range of developers' emotions during change tasks and are developers' emotions correlated with their perceived progress?

RQ2: What are aspects and practices that affect developers' emotions and progress during change tasks?

RQ3: Can we use biometric sensors to determine developers' emotions and progress during change tasks?

To address our research questions, we performed a study with 17 participants. In this study, participants worked on two change tasks for 30 minutes each while we recorded various biometric measures and periodically probed the participants for their emotions and progress. The results of our study show that developers experience a broad range of positive and negative emotions during change tasks that are similar to the ones experienced in other situations and that these emotions are highly correlated with progress, further supporting Graziotin *et al.*'s finding [9]. The results also show that the localization and understanding of relevant code are the most common aspects for emotions and progress to change. Using the biometric data gathered throughout the study, we trained a machine learning classifier that is able to distinguish between positive and negative emotions with an accuracy of 71.36% and between low and high progress with an accuracy of 67.70%.

This paper makes the following contributions:

- It presents and discusses the results of a study on the emotions and progress software developers have while working on change tasks.
- It presents an approach based on biometric measures to classify a developer's emotions and perceived progress during software development change tasks.

The results of our study suggest that we might be able to use biometric sensors to determine a developer's emotion and progress while working. This opens up a lot of opportunities for improving a developer's productivity.

II. Related Work

Work related to our research can be broadly categorized into four areas: general research on emotions and biometrics, research on biometric sensors in software engineering, research on developers' emotions and performance, and research on classifying progress.

A. Emotions and Biometrics

Research on emotions has a long history in psychology. Many theories and terminologies have been introduced along with several approaches to quantify emotions [17]. A widely used approach by Russel [18] differentiates between two cognitive dimensions of emotions: pleasure-displeasure and arousal-sleep. Today, these two dimensions are commonly called valence and arousal [19]. While the valence dimension is considered as the positive or negative character of an emotion [20], the arousal dimension indicates the amount of activation and excitement associated with an emotion [19]. In this paper, we generally adapt this terminology and refer to emotions with negative valence as negative emotions and emotions with positive valence as positive emotions.

To measure emotions, a broad range of research in psychology has explored the use of biometric sensors to measure the changes in the body caused by emotions. One of the most common emotions investigated through the use of biometric sensors is frustration. Researchers, for instance, induced frustration by manipulating computer games and measured the effect on the user with biometric sensors. Thereby, they found correlations between frustration and electro-dermal activity (EDA), blood volume pulse (BVP), electroencephalographic (EEG) activity, and muscle tension (*e.g.* [21]–[23]). In other studies, researchers found correlations between self-reported frustration levels and skin conductance or facial EMG while playing games or performing small tasks (*e.g.* [24], [25]).

To distinguish between various emotions, early research by Ekman *et al.* [26] was able to find differences in biometric signals for four negative emotions. More recently, similar studies were conducted that showed how BVP, EDA, respiration rate, or EEG can be used to distinguish between various emotions, such as anger, fear, sadness, disgust, happiness or surprise (*e.g.* [5], [27]–[29]).

Instead of distinguishing between different emotions, researchers have also used various biometric sensors to generally distinguish between positive and negative emotions. For instance, Leite *et al.* [30] used EDA to measure children's affective state while playing chess, finding that negative affective states are generally associated with an increased EDA signal that exhibits a lower variation. Reuderink *et al.* [22] found that EEG measures are correlated with the valence and arousal dimension when they studied subjects playing computer games and induced emotions through the use of non-responsive controllers. Muldner *et al.* [31] found that the pupil size changes with negative and positive affect when they studied subjects solving exercises in physics and varied the affect. Finally, Drachen *et al.* [32] used a combination of biometric measures while participants played a computer game and found that heart rate and EDA are correlated with self-reported negative/positive affect.

In our research, we built upon these previous findings, but focus on software developers performing realistic change tasks that stress a broad range of emotions and cognitive skills. Additionally, we investigate the use of such biometric sensors to predict progress.

B. Biometrics in Software Engineering

Only few studies in software engineering make use of biometric technology. Most of these focus on the use of eye tracking to examine program comprehension. For instance, Crosby *et al.* [33] and Bednarik *et al.* [13] used an eye tracker to study how experienced and less experienced developers understand source code. Similarly, Sharif *et al.* [15] relied on eye tracking technology to investigate how different identifier naming conventions influence program comprehension by examining the visual effort spent on identifiers.

Very few studies used other biometric sensors. Siegmund *et al.* [14] used functional magnetic resonance imaging (fMRI) to examine the brain regions that are activated during small program comprehension tasks. Parnin [34] investigated the use of electromyography to measure sub-vocal utterances and found that this could be used to determine programming task difficulty. Finally, in a previous study, we used a combination of biometric sensors and found that they can be used to assess the difficulty of small code comprehension tasks [16].

In contrast to these studies, we focus on the use of biometric sensors to classify developers' emotions and progress during change tasks.

C. Software Developers' Emotion & Progress

A few studies have investigated the emotions that software developers experience and how these emotions affect their progress and productivity. Early on, Shaw [8] observed 12 undergraduate students working on a software engineering project and found that the self-reported emotions can change drastically within 48 hours. Similarly, Wrobel *et al.* [1] conducted a survey to investigate how emotions impact software developers' effectiveness at work. They found that frustration is the most frequent negative emotion that also disturbs high productivity, and that for some people negative emotions can have a positive effect on productivity. Graziotin *et al.* [9] conducted an empirical study to investigate whether valence,

arousal and dominance correlate with the self-reported performance of software developers. In their study, they observed 8 developers working on a software development task for 90 minutes, asked them every 10 minutes about their emotions and productivity and found that valence and dominance are positively correlated with their productivity. In a second study, Graziotin *et al.* [35] observed 42 computer science students to find a relationship between affective states and creativity as well as analytic problem-solving skills. The study participants had to perform two tasks, an analytical one and a creative one, and affective states were assessed through a questionnaire. The results imply that developers with positive affect are significantly better problem solvers.

Different to these studies, Kahn *et al.* actively induced moods through videos that developers had to watch or influenced developers' arousal through physical exercises and found that developers' emotions have an influence on debugging performance [10]. Closer to our research, Khan *et al.* [12] also conducted two studies that focus on measuring mood with keyboard and mouse input. While in the first study, mood was self-reported, the second study induced mood through different kinds of music and an EDA sensor was used to differentiate between high and low arousal. The authors found an individual correlation between self-reported or induced mood and keyboard and mouse input, but no generic measure.

In our work, we extend results of earlier studies by providing more evidence on the correlation of emotions and progress, the aspects and practices that affect these and, in particular, how biometric measures can be used to classify self-reported emotions and progress during change tasks.

D. Classifying Progress

Research in psychology has shown that the state of being stuck and making no progress is frequently associated with negative emotions, while a state of flow and making lots of progress is frequently associated with positive emotions [3]. Only few approaches try to exploit this relationship and use biometric sensors to determine when people are in a state of being stuck or in flow. Muldner *et al.* [7], for instance, used four different sensors, a posture-chair, a skin conductance sensor, a pressure mouse, and an eye tracker, to determine the so called "*yes*-events"—brief expressions of positive affect—while students were solving physics exercises. They found that students had a larger pupillary response and a higher level of arousal during a yes-event, compared to neutral conditions. Our study is different in that we focus on realistic software change tasks and on emotions and progress.

In the field of software engineering, Carter *et al.* [11] tried to automatically determine moments in which a programmer is stuck based on IDE interaction logs and machine learning. In contrast to this work, we focus on biometric measures that are independent of an IDE and also differentiate between low and high progress as well as negative and positive emotions.

III. Study Method and Participants

To learn about developers' emotions and progress when performing change tasks and to address our research questions, we conducted a study with 17 developers. During the study, we had study participants work on two change tasks, while wearing biometric sensors. Additionally, we periodically asked them to assess their emotions and perceived progress.[1]

Fig. 1: Study setup with a subject in front of the eye tracker and computer screen, wearing the EEG headband and the Empatica wrist band.

A. Participants & Study Setup

For our study, we recruited 6 professional software developers and 11 PhD students with a major in computer science. Professional developers were recruited from two different software development companies in Switzerland. PhD students were recruited from the University of Zurich. The 17 participants (16 male, 1 female) ranged in age from 20 to 51 years and had an average professional development experience of 7.1 years ($\pm$ 6.7), ranging from 1 to 29 years.

Figure 1 depicts the study setup. For this study, we used three different sensors: an off-the-shelf Neurosky MindBand EEG sensor (http://neurosky.com/), an Empatica E3 wrist band (https://www.empatica.com/), and the Eye Tribe eye tracking device (https://theeyetribe.com/). The study took place in a quiet room. Study participants had to wear the EEG headband and the Empatica wrist band and were placed in front of a standard 1920 x 1080 24-inch screen with the eye tracker located in front of the screen.

B. Study Method

Subjects were first asked to put on the Empatica and the EEG sensors. We then ensured that the devices were connected, data was recorded properly and that the clocks of all recording devices were in sync. Before starting with the actual study, we instructed participants on the procedure and on how to rate their emotions and progress. Prior to starting on the first change task and before switching to the second change task, we asked participants to relax and watch a calming video of fish swimming in a fish tank for two minutes. In our previous study we saw that these two-minute videos relaxed participants and allowed their biometric features drop back

[1] A replication package of the study is available at http://seal.ifi.uzh.ch/people/mueller/SensingDevelopersEmotions.

TABLE I: Questions and answer ranges during our study.

1. Please rate how you felt right at the moment of the interruption. [-200 (very unpleasant) to +200 (very pleasant)]
2. Please rate how you felt right at the moment of the interruption. [-200 (very calm/relaxed) to +200 (very excited/stimulated)]
3. How do you rate your progress right before you were interrupted? [Likert scale with 1 (completely stuck / no progress at all), 3 (neutral), 5 (in flow / a lot of progress)]

to a baseline after about a minute [16]. After watching the fish tank video, participants started to work on one of the two change tasks. The order of the tasks was randomly assigned to each participant, but counterbalanced across all participants.

During their work on the change tasks, we interrupted participants either after they had been working for 5 minutes uninterrupted, or when they showed signs of strong negative or positive emotions, such as cursing or smiling. We chose a time frame of 5 minutes since previous studies found that developers switch tasks on average every 4.5 minutes during their work [36]. During each interruption, we asked participants to rate their emotional state at the moment of the interruption and their current perceived progress. For rating emotions, we followed Russell's 2-dimensional Circumplex model [18] and asked participants to rate them along two axes, a horizontal one for valence and a vertical one for arousal. Based on related work [37], both axes ranged from -200 (low) to +200 (high). To measure the perceived progress, we asked participants to rate it on a 5-point Likert scale. Table I lists the questions and answer ranges. In addition to these ratings, we asked participants about the reasons for their current state of emotions and progress and what could help them to feel better or make more progress. After working on the first change task for 30 minutes, we stopped participants, had them watch a fish tank video and then had them start working on the second task. For the second task we again followed the same protocol as for the first one.

After participants had been working on the second task for 30 minutes, we stopped them, showed them a two-minute fish tank video and then presented them two sets of pictures that are known for inducing specific emotions: one set inducing positive emotions and one inducing negative emotions [38]. The order of the two sets was randomly assigned to each participant, but counterbalanced across all participants. After each set of pictures, we asked participants to again rate their emotions. We used these picture sets to capture baselines of emotional reactions for each study participant. The picture sets were shown at the end of the study to ensure that they did not influence developers' emotions during their work on the change tasks. In between the two sets, we asked participants to relax and watch a fish tank video.

Once a study participant completed the last assessment of emotions, we stopped the recording of the biometric data and removed all sensors. Then we asked the participant to complete a questionnaire on the demographic background and conducted a brief interview. In the interviews we asked participants when and why they experience negative and positive emotions during change tasks and which practices they employ to avoid particularly negative emotions. We took hand written notes and audio recorded the interviews.

C. Change Tasks

Study participants were asked to work on two change tasks for which we provided short descriptions. One task was to write a small Java program that interacts with the StackExchange API [39] to retrieve all answers posted by a specific user on StackOverflow and sum up the scores the user earned for these answers. The other task was to implement a new feature in JHotDraw [40], an open-source Java GUI framework. JHotDraw provides a functionality to undo the latest command. For the study, participants were asked to implement a feature that allows users to undo more than one command at once by choosing from a history view of commands. We chose these two tasks, since they are representative of general change tasks as well as they are not too easy to solve and thus could stress both negative and positive emotions. We ran a pilot study with two subjects working on these two tasks, validating that they can trigger both positive and negative emotions. During the study, participants were allowed to use the Internet and search for help as they would normally do.

D. Data Collection

During the course of the study we collected biometric measurements that research has previously linked to negative and positive emotions: electro-dermal activity (EDA), electro-encephalography (EEG), skin temperature, heart rate, blood volume pulse (BVP) and various eye-related measurements, such as pupil size. We used a Neurosky MindBand sensor to capture EEG data, an Empatica E3 wrist band to record skin- and heart-related signals, and an Eye Tribe eye tracker to capture eye-related measures. Table II presents an overview of the captured measurements and the linked emotional aspects.

Over all study participants, we collected 213 emotion and 213 progress ratings with an average of 12.5 ($\pm$ 0.9) per participant, ranging from 11 to 14 ratings (as listed in the "Emotions" column of Table VII) and an interruption every 4.1 minutes on average, ranging from 0.7 to 5 minutes. In addition, we collected two emotion ratings per participant for the two sets of emotion inducing pictures. Finally, we also collected answers on the questions we asked for each of the 213 times we interrupted them and for the questionnaires and the interviews at the end.

IV. Analysis and Results

In this section, we address our three research questions by presenting the analysis and results of the quantitative and qualitative data gathered in our study. We present results aggregated over all participants. In a performance analysis between the group of professional developers and the group of PhD students we only found minor and statistically non significant differences. For the analysis, we divided each task into five subtasks and assessed for each participant the completion of each subtask as fully, partially or not at all completed. Professional developers fully or partially completed a mean

TABLE II: Overview of biometric measures and emotion-related aspects previously linked in literature.

Measure	Previously linked to
Eye-related	
Pupil size	excitement [7]; positive and negative affect [31]
Fixations	valence [41]
Brain-related	
Eye blinks	frustration [24]; stress and anxiety [42]
Frequency bands	valence [22], [43]; arousal [22]; happiness and sadness [44]; various emotions [28], [45]
Ratios of frequency bands	valence and arousal [46]
Attention and Meditation	valence and arousal [47]
Skin-related	
Electro-dermal activity (EDA)	valence and arousal [48], [49]; engagement [49]; frustration [23], [24], [29], [50]; positive and negative affect [30], [32]; various emotions [5], [26], [27], [51]
Skin temperature	valence and arousal [48]; boredom, engagement and anxiety [27]; various emotions [26]
Heart-related	
Blood volume pulse (BVP)	frustration [23]; various emotions [5]; valence and arousal [48];
Heart rate variability (HRV)	anxiety [52]; various emotional states [53]
Heart rate (HR)	valence [43], [48]; arousal [48]; positive and negative affect [32]; happiness [54]; various emotions [27], [51], [53]

of 6.5 of the 10 subtasks (3.5 fully, 3.0 partially), while PhD students fully or partially completed a mean of 6.4 subtasks (3.8 fully, 2.6 partially).

A. Experienced Range of Emotions & Progress (RQ1)

To examine the range of emotions developers experience during their work on a change task and whether the range of emotions is similar to the one experienced in other situations, we analyzed participants' ratings of arousal and valence during change tasks as well as the ones gathered for the emotion inducing pictures. Figure 2 illustrates the range of emotions with respect to the valence and arousal dimension. Blue markers indicate ratings during change tasks and red markers indicate ratings for the pictures that induce negative (low valence) and positive emotions (high valence). While working on the change tasks, each developer experienced a broad range of emotions, both for the valence and the arousal dimension. Valence ranged for developers from -183 to 200 and arousal from -151 to 181 with an average interquartile range per developer of 65.9 ($\pm$49.0) for the valence and 49.8 ($\pm$40.0) for the arousal dimension.

Since we are particularly interested in positive and negative emotions which are represented by the valence dimension, we compared the valence ratings for change tasks with the ones for the pictures inducing positive and negative emotions. Figure 3 depicts box plots of the valence ratings for change tasks as well as the two picture sets. The box plots highlight that the picture sets generally induced negative and positive emotions and that the range of emotions during change tasks and for the picture sets strongly overlap. These results show that *the emotions developers experience during change tasks cover a broad range of positive and negative emotions and that they are similar to the ones experienced in other situations.*

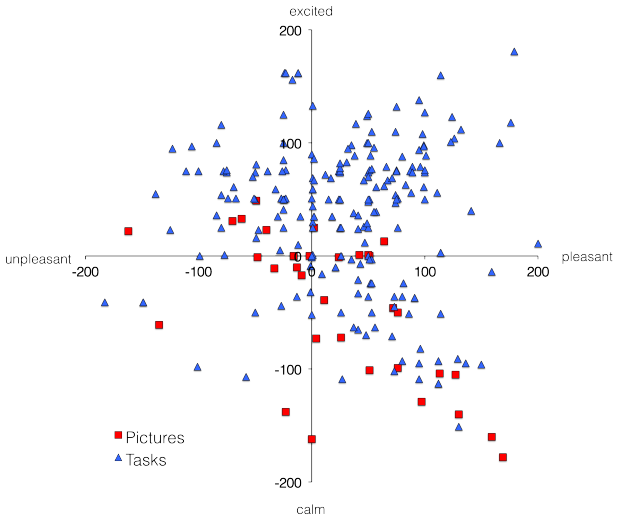


Fig. 2: Participants' emotion ratings on valence (x-axis) and arousal dimension (y-axis) during change tasks (▲) and after looking at emotion inducing pictures (■).

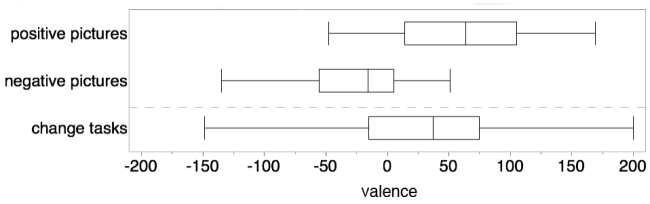


Fig. 3: Box plots of valence ratings after looking at positive / negative emotions inducing pictures, and during change tasks.

TABLE III: Progress ratings for the two change tasks.

	stuck		neutral		in flow
	1	2	3	4	5
Task 1	13	17	23	44	9
Task 2	18	31	29	28	1

During their work on each of the two change task, developers also experienced the whole range of progress, from 1 (being stuck) to 5 (in flow / a lot of progress), with a median of 3 and an interquartile range of 2 (see Table III). To examine whether emotion and progress ratings correlate and one might be able to use emotions as a proxy for a developer's progress, we applied a linear mixed model approach to the gathered data. We used a linear mixed model approach instead of other regression models, since research has shown that it is well suited for repeated measures from the same individual and is able to account for random effects, such as the task or the time of measurement [55]. We defined the self-reported progress rating as the dependent variable. Studies have shown that ordinal data, such as Likert scale ratings, can be used in these kind of parametric tests (*e.g.* [56]). Furthermore, we defined the valence and arousal as well as their interaction with the measurement time as fixed effects and the participant,

TABLE IV: Fixed-effects estimates on progress (* indicates significant estimates at the 0.05 confidence level).

Effect	Estimate	Upper p-value (207 df)	Lower p-value (181 df)	Deviance explained (%)
Valence	0.66 (*)	0.00	0.00	28.03
Arousal	0.10 (*)	0.02	0.02	1.09
Time	0.01	0.66	0.66	0.04
Valence:Time	-0.01	0.78	0.78	0.02
Arousal:Time	0.01	0.78	0.78	0.02

the task and the measurement time itself as random effects. Thereby, we standardized valence and arousal ratings for each participant to accommodate for individual differences in rating. Checking this model against the null model without any fixed effects results in a significant difference ($\chi^2(5) = 106.69, p < 0.001$). This difference shows that the valence and arousal dimensions have a significant effect on the progress in our model.

Table IV provides an overview of the fixed-effects estimates as well as the upper- and lower-bound p-values for assessing significance. The results show that, at the 0.05 confidence level, both arousal and valence are correlated with progress, however, the correlation between arousal and progress is only very weak. The valence dimension holds by far the highest explanatory power (28.03%) of the whole model (29.12%). The random effects for the measurement time is estimated to be 0, for the task it is in the range [-0.16, 0.16] and the random participant effect is in the range [-1.22, 0.80]. These results indicate that for the random effects that we modelled in our approach, the measurement time has no effect, the task has a medium effect and the participant has the highest effect on our model.

When analyzing individual ratings for each participant, we noticed that for some subjects the valence dimension of their emotions strongly correlated with the progress rating, while for other subjects, it did not. When calculating correlations on an individual basis, we found significant correlations for 12 subjects, but not for the other 5 subjects (S6, S7, S8, S14 and S16). Figure 4 provides examples for each of the two groups, with a strong correlation between valence and progress for subject S1 (Spearman's rank correlation coefficient $\rho = 0.88, p < 0.001$) in Figure 4a, and no significant correlation for subject S6 ($\rho = 0.38, p = 0.28$) in Figure 4b.

In summary, the results provide evidence that *valence is highly correlated with perceived progress* and might be a good indicator for progress overall, but a lot better for some individuals than others. These findings also support and confirm results on the correlation between emotion and progress found in a study with less participants by Graziotin *et al.* [9].

B. Aspects & Practices Affecting Emotions & Progress (RQ2)

To explore the aspects that affect emotions and progress during a change task and the practices developers employ to avoid negative emotions or a lack of progress, we analyzed the participants' answers to our questions during and after the tasks. We gathered a total of 186 answers out of the 213 data points in which participants' valence, progress or both ratings changed with respect to participants' previous rating during the change task. Since neither the participant's valence or progress changed for the other 27 data points, we did not include them in this analysis. In 91 of these 186 cases the valence and/or progress increased and in 95 cases they decreased. Based on grounded theory techniques [57], we used a combination of open, axial and selective coding to identify codes, group them into concepts and categories and find quotes related to the main categories. To avoid observer bias, both authors of this paper performed the coding, discussed and integrated the results.

TABLE V: Top 5 reasons for a change in emotions/progress.

Increase in emotions/progress	# Cases	# Subjects
localize relevant code	21 (11.3%)	14 (82.4%)
(better) understand parts of the code	18 (9.7%)	13 (76.5%)
next steps are clear	12 (6.5%)	9 (52.9%)
produce something / write code	9 (4.8%)	6 (35.3%)
have new idea	8 (4.3%)	6 (35.3%)
Decrease in emotions/progress	# Cases	# Subjects
difficulty in understanding how parts of the code/API work	33 (17.7%)	12 (70.6%)
difficulty in localizing relevant code	15 (8.1%)	8 (47.1%)
not being sure about next steps	9 (4.8%)	9 (52.9%)
realize that hypothesis on how code works is wrong	9 (4.8%)	7 (41.2%)
missing / insufficient documentation	3 (1.6%)	3 (17.6%)

Aspects that Affect Emotions & Progress. The top five reasons for an increase or a decrease in emotions or progress that participants mentioned when asked during their work on the tasks are summarized in Table V. In both cases, the ability to locate relevant parts of the code and the understanding of parts of the code are the top most reasons and account for 86 of the total of 186 cases (46.2%). When participants were able to locate a starting point or relevant code, it made them feel better and have a good feeling of their progress, while not finding the relevant code resulted in the opposite:

"I've found a starting point almost immediately. I have a good feeling that I'll make significant progress very soon." (S1)

"It's going too slow. I think it's very cumbersome when so much time is needed to understand the project and find a starting point." (S9)

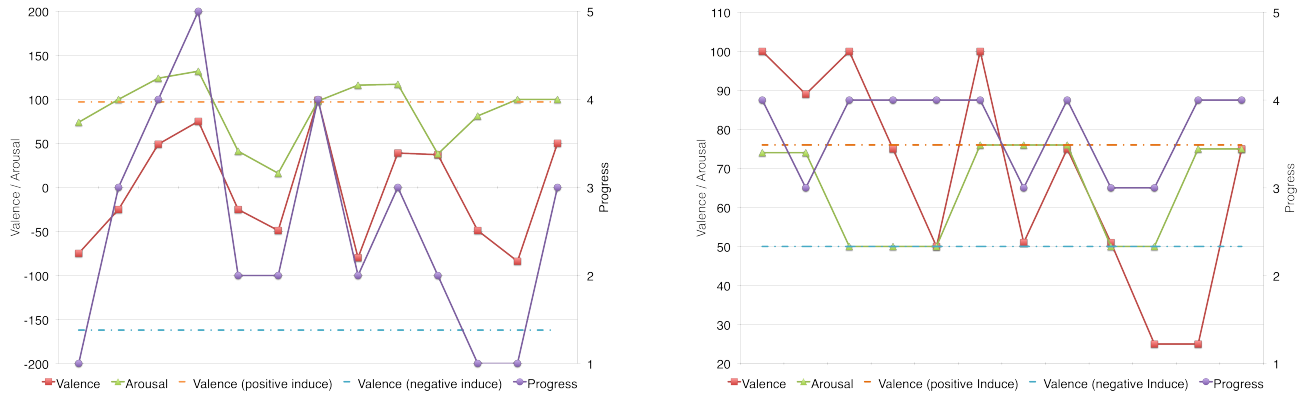
Similarly, a better understanding of parts of the code or the difficulty in understanding can cause changes in a developer's emotions and her feeling of progress. While a better understanding of parts of the code can lift a developer's emotions, the lack thereof can lead to annoyance and anger:

"I finally understand what I really need to do. There is light at the end of the tunnel." (S12)

"It's unclear how to use `UndoManager`*. That pisses me off." (S1)*

One aspect that had a very strong impact on a developer's emotion or perception of progress was the writing of code. Thereby, it was not even important whether the code is correct or not, just the mere fact of producing some visible output lifts the spirit of developers as one participant stated:

"I'm feeling slightly better again, since I've produced something visible. At the moment, it's not so important if it's correct or not.



(a) S1 (b) S6

Fig. 4: Emotion and progress ratings for S1 and S6.

Most importantly, I've produced output. That makes me feel great." (S7)

Common among all answers is that participants talked about emotions and progress simultaneously. In many cases, emotions were either mentioned alongside with a perception of progress, or they were mentioned as one affecting the other, for instance, the lack of progress causing annoyance:

"I finally figured out how to do it. I'm really happy and I'm not feeling completely stuck anymore." (S6)

"I can't make any progress. That's annoying." (S13)

This co-occurrence of comments on emotions and on progress in participants' answers further supports our findings from RQ1, indicating a correlation between a developer's positive and negative emotions and the perceived progress.

Practices Employed to Avoid Negative Emotions and Getting Stuck. Since we are interested in understanding how we can support developers in avoiding negative emotions and being stuck, we also asked about the support one could provide in these cases. Most commonly participants stated that a more complete and detailed documentation (27 cases), a description of the high-level architecture (18 cases) and better code examples (17 cases) would be beneficial to feel better and improve progress. When asked more generally about the practices participants employ to avoid negative emotions, three general strategies emerged from the answers: *switching context when stuck, setting clear goals, and allocating sufficient resources ahead of time*.

Several participants stated that they will switch context and, for example, switch to a different task, talk to others, or take a break. This helps them to feel better and to get new ideas:

"When I'm frustrated, most often I take a coffee break or do something completely different. For example read the [news] online. Just something completely different. Most often I'm more relaxed afterwards." (S9)

"I take a break then and suddenly after the break, the problem is way easier to solve." (S2)

Another practice to avoid negative feelings in the first place, is to set yourself clear goals before starting to work and then actively avoid potential distractions and fade out all other things. Participants thereby also mentioned to give themselves certain rewards for achieving a certain goal:

"What helps me is to set myself a goal. For example, work on this task until then and then, and afterwards, I will give myself some sort of reward, for example, take a break." (S1)

Finally, allocating and planning sufficient resources for a task is a common strategy among participants to avoid negative emotions. Study participants reported that time pressure often leads to stress and frustration for them and they therefore try to reserve enough time for the completion of a task.

C. Biometric Sensors to Determine Developers' Emotions and Progress (RQ3)

To investigate whether we can use biometric sensors to distinguish between positive and negative emotions as well as episodes of low and high progress that developers experience during change tasks, we applied a machine learning approach to the collected data. Over the course of the participants' work on both change tasks we collected biometric data for a total of 213 intervals. Figure 5 illustrates a set of four such intervals for participant S4 together with the collected EDA and the heart rate signal as well as the participant's emotion and progress ratings. Especially for the EDA signal, the example presented in Figure 5 shows a visible difference between the first episode with medium progress and higher valence compared to the last episode with the developer being stuck and a lower valence. Each interval is delimited by our periodic interruptions for which the data is not taken into account. Since emotions typically last for seconds or at the longest up to minutes [58], we decided to take into account the 10 seconds of biometric data collected before each time we asked a participant to rate her emotions and progress. Due to errors during the data capturing process, we were not able to collect or use heart- and skin-related measurements for S3 and eye-related measurements for S8 in our analysis. All other measurements for these two participants were, however, included in the analysis.

Data Cleaning and Feature Extraction. Since biometric data can be intensely noisy, we applied various noise cleaning steps to the data before extracting features. For the eye-tracking

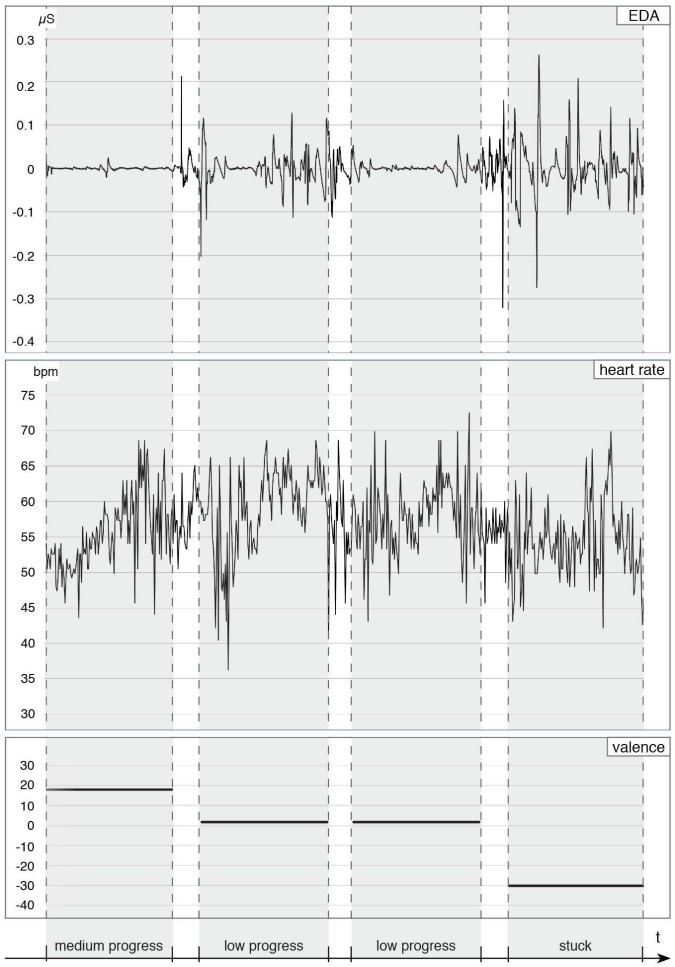


Fig. 5: Exemplary biometric data, emotion and progress ratings collected over 4 intervals for participant S4.

data, we deleted all measurements that were marked as invalid by the eye-tracking device. We also standardized the pupil sizes by participant to account for the differences between participants. Based on research that has shown that pupil size and fixation duration is affected by positive and negative emotions (*e.g.* [7], [31], [41]), we extracted various features for fixation duration and pupil size.

The EEG sensor captures the electrical activity of the brain, measured on the scalp. Research has shown that frequency bands, extracted from brain waves, can be used to distinguish between various emotions (*e.g.* [28], [44], [45]). We therefore extracted common frequency bands [59]: α (8-12Hz), β (12-30Hz), γ (30-80Hz), δ (0-4Hz), θ (4-8Hz) from the brain wave signal and also calculated the fraction of each band with one another. Additionally, the Neurosky Mindband sensor provides two pre-processed signals, called Attention and Meditation, that we also used for our analysis. Finally, we extracted the eye blink rate from the EEG signal using a method proposed by Manoilov [60]. We could have extracted the eye blink rate from the eye tracker data. However, since we were not able to capture this data for S8, we used the EEG signal.

The EDA signal consists of two parts: the low frequency, slowly changing tonic part, and the high frequency, fast changing phasic part [61]. We used a low-pass and a high-pass Butterworth filter to extract the phasic and tonic part from the EDA signal. In particular features related to the peaks in the phasic signal, but also features extracted from the tonic part of the signal, were closely linked to emotions in previous studies (*e.g.* [5], [30], [32]). The Empatica E3 sensor also measures skin temperature that research has used to infer emotional states (*e.g.* [26], [27], [48]). We included these features in our analysis as well.

For the heart-related data we focused on features that describe peaks of the blood volume pulse (BVP) signal [62] and we also extracted the mean heart rate that was used in research to asses emotions [43], [54]. By a simple transformation, the heart rate can be used to calculate the heart rate variability (HRV) that represents the variation in the time interval between two consecutive heart beats. In research, HRV was previously used to infer various emotional states [63] and is usually analyzed by calculating the mean and the standard deviation of the time between two successive heart beats [64]. We added these features to our analysis as well.

Data Labelling. To distinguish between positive and negative emotions, we focused on the valence dimension of the emotion ratings. Given the individual differences in the way participants rated their emotions, we used the emotion inducing pictures, known for inducing particularly negative and positive emotions [38], to determine a baseline for each participant. We calculated the mean of the valence ratings for the positive and negative emotion inducing pictures and then labelled the ratings from the change tasks below the mean as negative and the ones above the mean as positive. We manually inspected all valence ratings for each participant to disambiguate in cases the ratings were very close to the mean. Only for 5 cases out of the 213, the ratings were almost identical with the mean. In these five cases we additionally took the participants' comments into account to unambiguously label them. We ended up with 128 ratings for positive and 85 ratings for negative emotions.

To distinguish between low and high progress, we used participants progress ratings on the 5-point Likert scale (1 = completely stuck / no progress at all, 3 = neutral, 5 = in flow / a lot of progress) and classified ratings of 1 and 2 as low progress and ratings of 4 and 5 as high progress. Since we were not interested in episodes where subjects reported their progress as neutral, we removed these instances from our analysis. We ended up with 79 instances of low progress and 82 instances of high progress.

Machine Learning. For our machine learning classifier, we used the Java-based framework Weka [65]. We first partitioned our data by participant and task. Since each participant worked on two different change tasks and we collected 5 to 7 emotion and progress ratings per task and participant, we ended up with 17 times 2 (=34) participant-task combinations, each having 5 to 7 data points. We partitioned the data by participant and task to avoid having data points from the same participant and the same task in both the training and testing set. We then

used a leave-one-out method and trained our classifier in turn with all participant-task combinations except one, and used the remaining combination as test set. For feature selection we used ConsistencySubsetEval, a Weka implementation of an algorithm that chooses a feature subset based on the consistency between the data [66]. As classifier, we opted for a decision tree classifier and used J48, the Weka implementation of C4.5 [67]. We used a decision tree classifier under the assumption that its non-parametric characteristics [68] would fit our collected data, which often exhibited a non-parametric distribution.

Results. Table VI presents the results of our machine learning classification. When classifying emotions into positive and negative ones, a classifier trained on biometric data is able to predict 71.36% of all cases correctly. Compared to a naive predictor that always predicts the most dominant class but never any other class, this is an improvement of 18.76%. Compared to a random predictor that randomly predicts one of the two classes, this is an improvement of 42.72%. The features with the most predictive power for this kind of prediction were the brainwave frequency bands, the pupil size, as well as the heart rate. Predicting the progress achieved similar accuracy. Our machine learning approach was able to classify 109 out of 161 cases correctly (67.70%). This is an improvement of 32.93% compared to a naive predictor and 35.40% compared to a random predictor. To classify progress, the EDA tonic signal, the temperature, brainwave frequency bands, and the pupil size were most predictive. These results indicate that we can use biometric measurements to distinguish fairly accurately between positive and negative emotions that a developer experiences during a programming task. Slightly less accurate, but still better than a naive or random predictor, it is also possible to distinguish between episodes of low and high perceived progress. The results also indicate that a combination of multiple sensors works best for these kind of predictions.

To examine the results in more detail and by participant, Table VII lists all results partitioned by participant. The results show that for some participants, *e.g.* S11 or S14, the prediction of positive and negative emotions as well as low and high progress works very well, while for other participants, such as S16 or S17, both predictions do not achieve great accuracy.

Finally, we also examined if we could train a classifier and then use it for classifying emotions and progress of a participant that the classifier was not previously trained on. Therefore, we trained the classifier in turn for all participants except one and used the data of the remaining participant for testing. While the accuracy for distinguishing positive and negative emotions is identical (71.36%), the accuracy to distinguish between low and high progress is slightly lower (63.35%).

TABLE VI: Machine learning results for classifying emotions and progress together with the features selected for each classifier (Δ represents the difference to the baseline).

Prediction	Correct	Precision	Recall	Selected features
Emotion	71.36%	64.32%	82.03%	ΔAlpha, ΔBeta/Theta MinPupilSize, ΔMeanHR
Progress	67.70%	67.85%	68.29%	ΔAlpha, ΔBeta/Theta ΔMeanTempPeakAmpl, MaxPupilSize, ΔMeanPupilSize, ΔMeanSCL

TABLE VII: Machine learning results partitioned by participant.

Participant	Emotions			Progress		
	Correct	Total	% Correct	Correct	Total	% Correct
S1	8	13	61.54	3	10	30.00
S2	11	14	78.57	6	10	60.00
S3	12	14	85.71	7	12	58.33
S4	8	11	72.73	6	7	85.71
S5	11	13	84.62	6	10	60.00
S6	8	11	72.73	4	5	80.00
S7	4	13	30.77	8	10	80.00
S8	10	13	76.92	4	7	57.14
S9	11	13	84.62	6	9	66.67
S10	9	13	69.23	8	11	72.73
S11	10	12	83.33	9	9	100.00
S12	10	12	83.33	4	9	44.44
S13	10	13	76.92	6	10	60.00
S14	10	12	83.33	9	12	75.00
S15	6	12	50.00	11	11	100.00
S16	7	12	58.33	5	8	62.50
S17	7	12	58.33	7	11	63.64
Total	**152**	**213**	**71.36**	**109**	**161**	**67.70**

V. Discussion

Individual Differences. While our results show that over all developers there is a correlation between emotions and perceived progress and biometric features can be used to predict emotions and progress, the analysis also shows that there are strong individual differences with respect to the correlation and the classification. For instance, while the machine learning classifier for emotions is only correct in 30.77% of the cases for participant S7, it goes up to 85.71% of correct cases for S3 (see Table VII). Khan *et al.* [12] already pointed out that, due to the widely varying range and perception of emotions by individual participants, it is very difficult to find relationships between an individual's emotions and other aspects, such as progress, that hold for more than just a few people. Especially given the high variability in biometric measures between individuals [69], an approach that more specifically takes into account individual differences and is, for instance, trained specifically for a developer, might greatly improve our results. While this has the disadvantage of requiring training sessions for each user, we assume this will be negligible compared to the potential benefit it could bring.

Developer Support. The quantitative and qualitative results of our study also contribute new knowledge on how to support a developer and ensure that her time is spent as productive as possible. Since developers frequently had negative emotions and experienced little progress when they did not understand parts of the code, one could use an approach with a classifier on emotions and progress to identify places in the code that are particularly difficult to understand and might benefit from a code review or refactoring. In particular, given the new advances in eye tracking technology that make them a lot more affordable and easy to add to any existing setup, one might be able to use the classifier to track the difficult parts of the code on the level of code lines. Thus, such a classifier might be used as a new kind of code smell detector

that could more automatically add a human aspect to code analysis. Furthermore, when a developer is trying to locate relevant code, a classifier on emotions and progress could be used to determine the times when code recommendations would be particularly helpful. This would allow to avoid overwhelming developers with continuous recommendations but provide them at opportune moments when the developer is most susceptible to them. In addition, knowing when a developer has particularly negative emotions or getting stuck one could recommend taking a break, while in a state of flow, tool support could be built to avoid interruptions through notifications or coworkers.

Negative Emotions. While results have shown that negative emotions are often correlated with low progress, it is important to note that avoiding negative emotions at every cost will not always automatically lead to more progress. In certain situations, negative emotions might actually be a necessary part to solve a problem and lead to higher progress later on, and occasionally being frustrated by a task might provide indirect benefits. As Wrobel *et al.* [1] already observed in their study, negative emotions can act as an activator for developers to become more productive. Future studies are needed to explore this positive aspect of negative emotions further and determine if there are ways to distinguish between possibly beneficial or detrimental experiences of negative emotions, such as frustration.

Ethical and Privacy Concerns. Finally, with the introduction of biometric sensors there are also ethical and privacy concerns to be addressed. While advances in sensor technology might decrease the physical invasiveness of these sensors, the capturing of huge amounts of very personal data can raise several ethical and privacy concerns in a developer. By focusing on the individual and providing personalized support to the developer without sharing the fine-grained and very developer-specific data with others, we can help to assuage these concerns. In future studies, we plan to further investigate the impact of such support on the individual and how we might be able to avoid such concerns.

VI. Threats To Validity

External Validity. Since participants only worked on two change tasks, the generalizability of our study might be limited. We tried to mitigate this risk by carefully choosing study tasks representative of typical change tasks, either requiring the use of a popular API or requesting a change in a system commonly used for studies. Another threat to generalizability is the selection of participants. We tried to limit this by recruiting participants with various backgrounds.

Internal Validity. We observed participants while they were working in our lab study setup. In particular the environment might trigger different emotions or progress than participants would usually experience in their work environment. We tried to mitigate this risk by selecting representative tasks that triggered a broad range of emotions and validated it with emotion inducing pictures. Future work needs to investigate if these results can be ported to life work environments.

During the study, we regularly interrupted participants and asked them to rate their emotions and progress. These interruptions might have influenced participants' performance and ratings. We tried to mitigate this risk by choosing a time interval between interruptions that is representative of the time interval of developers' task switches.

Construct Validity. As one part of this study, we used biometric measurements to predict the positive and negative emotions as well as developers' perceived progress. The data captured with these sensors might be affected not only by the emotions and progress that developers experienced during the study tasks, but also by study participants' personality traits or their general stress level. To mitigate this risk, we conducted the study in a quiet environment and limited all unnecessary distractions. Additionally, we periodically let the study participants watch a calming and relaxing video and collected biometric baseline data that we used in the analysis to compare the data collected during the study tasks.

VII. Conclusion

Software developers experience a broad range of emotions during their work. Previous studies have shown that biometric sensors can be used to distinguish between positive and negative emotions. These studies focused on certain kinds of tasks, such as very small analytical tasks, that are not representative of development tasks. In the presented research, we built upon and extend previous work to the context of software change tasks and the classification of a developer's progress as well as emotions. The results of our study show that using machine learning, we are able to distinguish between positive and negative emotions in 71.36% of all cases and between low and high progress in 67.70%. Our results also show that emotions and perceived progress are highly correlated and illustrate aspects and practices that affect emotions and progress. These insights provide a lot of opportunities for future work that could have direct potential impact on a developer's work and productivity. One could, for instance, provide automatic support to a developer by recommending code examples, relevant documentation or even just a short break when the developer is getting stuck and frustrated and it might be most beneficial. Similarly, one can imagine tool support to minimize and postpone interruptions when the developer is in flow and the cost of an interruption would be particularly high. In future work, we intend to explore these opportunities and further examine more individualized classifiers that take into account differences in people's biometric data and thus might be able to determine a developer's emotions and progress even better.

Acknowledgment

We thank all subjects for participating in our study and Elaine Huang, Gail Murphy and the reviewers for their valuable feedback.

REFERENCES

[1] M. Wrobel, "Emotions in the software development process," in *International Conference on Human System Interaction (HSI)*, 2013, pp. 518–523.

[2] A. P. Brief and H. M. Weiss, "Organizational behavior: affect in the workplace." *Annual Review of Psychology*, vol. 53, pp. 279–307, 2002.

[3] W. Burleson and R. Picard, "Affective agents: Sustaining motivation to learn through failure and a state of stuck," in *Workshop on Social and Emotional Intelligence in Learning Environments*, 2004.

[4] R. Lawson, *Frustration: The Development of a Scientitifc Concept*. The Macmillan Company, 1965.

[5] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1109/34.954607

[6] B. Reuderink, A. Nijholt, and M. Poel, "Affective pacman: A frustrating game for brain-computer interface experiments," in *Intelligent Technologies for Interactive Entertainment*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, A. Nijholt, D. Reidsma, and H. Hondorp, Eds. Springer Berlin Heidelberg, 5 2009.

[7] K. Muldner, W. Burleson, and K. VanLehn, ""Yes!": using tutor and sensor data to predict moments of delight during instructional activities," in *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, ser. UMAP'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 159–170.

[8] T. Shaw, "The emotions of systems developers: An empirical study of affective events theory," in *Proceedings of the 2004 SIGMIS Conference on Computer Personnel Research: Careers, Culture, and Ethics in a Networked Environment*, ser. SIGMIS CPR '04. New York, NY, USA: ACM, 2004, pp. 124–126. [Online]. Available: http://doi.acm.org/10.1145/982372.982403

[9] X. W. Daniel Graziotin and P. Abrahamsson, "Are happy developers more productive? the correlation of affective states of software developers ane their-self-assessed productivity," in *Proceedings of the 14th Internationl Conference on Product-Focused Software Process Improvement*, 2013.

[10] I. A. Khan, W.-P. Brinkman, and R. M. Hierons, "Do moods affect programmers' debug performance?" *Cognition, Technology & Work*, vol. 13, no. 4, pp. 245–258, 2011.

[11] J. Carter and P. Dewan, "Design, implementation, and evaluation of an approach for determining when programmers are having difficulty," in *Proceedings of the 16th ACM International Conference on Supporting Group Work*, ser. GROUP '10. New York, NY, USA: ACM, 2010, pp. 215–224. [Online]. Available: http://doi.acm.org/10.1145/1880071.1880109

[12] R. H. Iftikhar Ahmed KHAN, Willem-Paul Brinkman, "Towards estimating computer users' mood from interaction behaviour with keyboard and mouse," *Frontiers of Computer Science*, 2013.

[13] R. Bednarik and M. Tukiainen, "An eye-tracking methodology for characterizing program comprehension processes," in *Proceedings of the Symposium on Eye Tracking Research & Applications*. ACM, 2006, pp. 125–132.

[14] J. Siegmund, C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, and A. Brechmann, "Understanding understanding source code with functional magnetic resonance imaging," in *Proceedings of the 36th International Conference on Software Engineering*. New York, NY, USA: ACM, 2014, pp. 378–389. [Online]. Available: http://doi.acm.org/10.1145/2568225.2568252

[15] B. Sharif and J. I. Maletic, "An eye tracking study on camelcase and under_score identifier styles," in *18th International Conference on Program Comprehension (ICPC)*. IEEE, 2010, pp. 196–205.

[16] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 402–413. [Online]. Available: http://doi.acm.org/10.1145/2568225.2568266

[17] P. Ekkekakis, *Measurement in sport and exercise psychology*. Human Kinetics, 2012, ch. Affect, Mood, and Emotion.

[18] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[19] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological Review*, vol. 110, no. 1, pp. 145–172, Jan 2003.

[20] G. Colombetti, "Appraising valence," *Journal of consciousness studies*, vol. 12, no. 8-10, pp. 103–126, Aug. 2005.

[21] J. Riseberg, J. Klein, R. Fernandez, and R. W. Picard, "Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state," in *Conference Summary on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1998, pp. 227–228. [Online]. Available: http://doi.acm.org/10.1145/286498.286715

[22] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the eeg during game play," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 6, no. 1, pp. 45–62, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1504/IJAACS.2013.050691

[23] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard, "Frustrating the user on purpose: a step toward building an affective computer," *Interacting with Computers*, vol. 14, no. 2, pp. 93 – 118, 2002.

[24] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2007.02.003

[25] R. Hazlett, "Measurement of user frustration: A biologic approach," in *Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2003, pp. 734–735. [Online]. Available: http://doi.acm.org/10.1145/765891.765958

[26] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions." *Science*, vol. 221, no. 4616, pp. 1208–1210, Sep 1983.

[27] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*. New York, NY, USA: ACM, 2008, pp. 13–17. [Online]. Available: http://doi.acm.org/10.1145/1457199.1457203

[28] M. Murugappan, M. Rizon, R. Nagarajan, and S. Yaacob, "Eeg feature extraction for classifying emotions using fcm and fkm," in *Proceedings of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science*, ser. ACACOS'08. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2008, pp. 299–304. [Online]. Available: http://dl.acm.org/citation.cfm?id=1415743.1415793

[29] R. Mandryk, K. Inkpen, and T. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Special Issue on User Experience, Behaviour and Information Technology*, vol. 25, no. 2, pp. 141–158, 2006.

[30] I. Leite, R. Henriques, C. Martinho, and A. Paiva, "Sensors in the wild: Exploring electrodermal activity in child-robot interaction," in *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2013, pp. 41–48.

[31] K. Muldner, R. Christopherson, R. Atkinson, and W. Burleson, "Investigating the utility of eye-tracking information on affect and reasoning for user modeling," in *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 138–149.

[32] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. New York, NY, USA: ACM, 2010, pp. 49–54. [Online]. Available: http://doi.acm.org/10.1145/1836135.1836143

[33] M. Crosby and J. Stelovsky, "How do we read algorithms? a case study," *Computer*, vol. 23, no. 1, pp. 25–35, 1990.

[34] C. Parnin, "Subvocalization-toward hearing the inner thoughts of developers," in *International Conference on Program Comprehension (ICPC)*. IEEE, 2011, pp. 197–200.

[35] D. Graziotin, X. Wang, and P. Abrahamsson, "Happy software developers solve problems better: psychological measurements in empirical software engineering." *PeerJ*, vol. 2, 2014.

[36] A. N. Meyer, T. Fritz, G. C. Murphy, and T. Zimmermann, "Software developers' perceptions of productivity," in *Proceedings of the International Symposium on the Foundations of Software Engineering*, 2014.

[37] G. Mark, S. T. Iqbal, M. Czerwinski, and P. Johns, "Bored mondays and focused afternoons: The rhythm of attention and online activity in the workplace," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New

York, NY, USA: ACM, 2014, pp. 3025–3034. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557204

[38] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance." *Behavior research methods*, vol. 43, no. 2, pp. 468–477, Jun 2011.

[39] "http://api.stackexchange.com/."

[40] "http://www.jhotdraw.org/."

[41] E. Carniglia, M. Caputi, V. Manfredi, D. Zambarbieri, and P. E., "The influence of emotional picture thematic content on exploratory eye movements," *Journal of Eye Movement Research*, 2012.

[42] D. G. Doehring, "The relation between manifest anxiety and rate of eyeblink in a stress situation," Central Institute for the Deaf, Tech. Rep., 1957.

[43] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, "Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music." *Psychophysiology*, vol. 44, no. 2, pp. 293–304, Mar 2007.

[44] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG." *Conference Proceedings Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2009, pp. 1323–1326, 2009.

[45] M. Murugappan, R. Nagarajan, and S. Yaacob, "Modified energy based time-frequency features for classifying human emotions using eeg," in *Proceedings of the International Conference on Man-Machine Systems*, 2009.

[46] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, July 2010.

[47] H. Yoon, S. wook Park, Y.-K. Lee, and J.-H. Jang, "Emotion recognition of serious game players using a simple brain computer interface," in *International Conference on ICT Convergence (ICTC)*, Oct 2013, pp. 783–786.

[48] A. H. S. G. P. S. J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Affective Dialogue Systems Lecture Notes in Computer Science*, vol. 3068, pp. 36–48, 2004.

[49] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: An intelligent system for emotional memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 849–858. [Online]. Available: http://doi.acm.org/10.1145/2207676.2208525

[50] G. L. Freeman, "A method of inducing frustration in human subjects and its influence upon palmar skin resistance," *The American Journal of Psychology*, vol. 53, no. 1, pp. pp. 117–120, 1940.

[51] J. Wagner, J. Kim, and E. Andre, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proceedings of the International Conference on Multimedia & Expo (ICME 2005)*, 2005, pp. 940–943.

[52] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system - towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1017/S0263574703005319

[53] R. McCraty and D. Tomasino, *Stress in Health and Deseases*. Wiley-VCH, 2006, ch. Emotional Stress, Positive Emotions, and Psychophysiological Coherence.

[54] A. Steptoe, J. Wardle, and M. Marmot, "Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes." *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6508–6512, May 2005.

[55] R. Gueorguieva and J. H. Krystal, "Move over anova: Progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry," *Archives of General Psychiatry*, vol. 61, no. 3, pp. 310–317, 2004. [Online]. Available: http://dx.doi.org/10.1001/archpsyc.61.3.310

[56] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Science Education: Theory and Practice*, vol. 15, no. 5, pp. 625–632, Dec 2010.

[57] P. Y. Martin and B. A. Turner, "Grounded theory and organizational research," *The Journal of Applied Behavioral Science*, 1986.

[58] P. Ekman, *The nature of emotion*. Oxford Univserity Press, 1994, ch. Moods, emotions, and traits.

[59] T. C. Handy, *Event-related potentials: a methods handbook*. MIT Press, 2005.

[60] P. Manoilov, "Eye-blinking artefacts analysis," in *Proceedings of the International Conference on Computer Systems and Technologies*, 2007, p. 52.

[61] S. Schmidt and H. Walach, "Electrodermal activity (EDA) - state-of-the-art measurements and techniques for parapsychological purposes," *Journal of Parapsychology*, vol. 64, no. 2, p. 139, June 2000.

[62] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability respiratory sinus arrhythmia, and cardiorespiratory synchrony?" *Biofeedback*, vol. 35, no. 2, pp. 54–61, 2007.

[63] J. E. Mietus, C.-K. Peng, I. Henry, R. L. Goldsmith, and A. L. Goldberger, "The pnnx files: re-examining a widely used heart rate variability measure." *Heart*, vol. 88, no. 4, pp. 378–380, Oct 2002.

[64] R. D. B. Kenneth C. Bilchick, "Heart rate variability," *Journal of Cardiovascular Electrophysiology*, 2006.

[65] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[66] H. Liu and R. Setiono, "A probabilistic approach to feature selection - a filter solution," in *13th International Conference on Machine Learning*, 1996, pp. 319–327.

[67] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[68] L. Rokach and O. Maimon, *Data Mining and Knowledge Discovery Handbook*. Oded Maimon and Lior Rokach, 2006, ch. Decision Trees.

[69] R. Mandryk, *Game Usability*. CRC Press, 2008, ch. Physiological Measures for Game Evaluation.